Beyond Tokens and Into Minds: Future Directions for Human-Centered Evaluation in Machine Translation Post-Editing

Molly Apsel[1]*, Sunil Kothari[2], Manish Mehta[2], and Vasudevan Sundarababu[2]

[1]Indiana University, [2]Centific

Correspondence: mapsel@iu.edu, manish.mehta@centific.com

Abstract

Machine translation post-editing (MTPE) is central to evaluating and ensuring translation quality, particularly for low-resource languages (LRLs), where systems are more error-prone than for high-resource languages. Traditional token-based models segment text according to statistical patterns of their (primarily high-resource) training data, which can distort meaning, fragment words in morphologically rich languages, and complicate MTPE and evaluation. Current evaluation metrics also tend to emphasize surface-level similarity to reference texts, overlooking how humans actually approach translation tasks and creating issues when references are unavailable or a more abstract interpretation is needed. In this position paper, we argue that emerging architectures (Large Concept Models [LCMs] and Byte Latent Transformers [BLTs]) and insights from cognitive science open new possibilities for MTPE frameworks. LCMs represent meaning at the conceptual level, enabling evaluation of different translation approaches and the robustness of such models in MT. At the same time, BLTs operate below the token level, potentially easing post-editing across diverse language scripts. Drawing on cognitive theories of bilingualism and meaning representation, we outline hypotheses and research methods for evaluating post-editing data, translation quality, and interface design toward more robust, human-centered MT evaluation.

1 Introduction

Machine translation post-editing (MTPE) has become a critical tool for ensuring the quality of machine translation. Post-editing involves human translators correcting machine outputs, which not only speeds up the overall translation process compared to manual translation alone but also provides feedback that can improve future MT quality.

However, the efficiency gains from an MTPE workflow can vary widely depending on several factors. First, the initial quality of the MT affects the effort required by post-editors. While MT systems have continued to evolve, especially with the advent of Transformer models, their success is often constrained by the amount of training data available in the source and target languages. This means that low-resource languages (LRLs), or languages that have limited digital language data or tools available (e.g., Swahili, Sinhala, Basque), are more likely to have severe translation errors, which require more effort on the part of post-editors (Haddow et al., 2022). Additionally, languages vary in syntactic structure and morphological richness, which is the amount of grammatical information expressed in each word. Language pairs with vastly different linguistic and morphological features are more cognitively demanding for post-editors. Because LRLs are less likely to have tokenizers that capture their linguistic structures, this challenge is often exacerbated for LRLs.

Further, MT performance is often assessed using automated metrics that compare outputs with reference translations, such as Bilingual Evaluation Understudy (BLEU). As a result, reported quality depends heavily on the reliability of these metrics and the availability of strong reference translations. The validity of these assessments can also vary significantly across language pairs. For instance, LRLs tend to have fewer reference translations available, and measures such as the number of edits might not accurately reflect the quality of the MT. While MTPE has become a valuable step toward enabling broader access to reliable translations, there is a vast opportunity to create systems that allow speakers of all languages to enjoy the potential benefits of MT and MTPE.

Recent advances in language modeling and research in cognitive science offer insights into how we might innovate MT workflows to address ex-

[*] Work done while interning at Centific.

Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems, pages 1–8
December 23, 2025 ©2025 Association for Computational Linguistics

isting gaps, especially for LRLs. Traditional MT models and evaluation metrics operate at the token level, which can impose limitations depending on the language pair and translation purpose. In 2024, Meta introduced two alternatives to token-based language models (LMs): the Byte Latent Transformer (BLT; Pagnoni et al., 2024) and the Large Concept Model (LCM; Barrault et al., 2024). While both move beyond fixed tokenization, they do so in contrasting ways – one by breaking text into finer-grained units, the other by abstracting above the level of text altogether.

The BLT operates at a more granular level, dynamically segmenting the input byte stream into variable-length units based on predictability and compression efficiency, allowing the model to adapt its representations rather than relying on a fixed tokenizer. This design not only improves computational efficiency but also reduces biases introduced by tokenizers that privilege dominant-language vocabularies. Pagnoni et al. demonstrated that BLT outperforms the Llama 3 token-based model on LRL translation both to English from other languages and vice versa.

On the other hand, the LCM aims to overcome the limitations of tokens by instead representing meaning at the level of abstract "concepts." These semantic representations are intended to be language- and modality-agnostic, so they are not tied to any particular language or information format. This approach promises universal, cross-lingual representations that capture the abstract ideas underlying a text rather than predicting one sequence of tokens from another, which may be more difficult to do across specific language pairs. The researchers who developed the LCM showed that it surpasses a Llama 3 model in a text summarization task for several LRLs (Barrault et al., 2024).

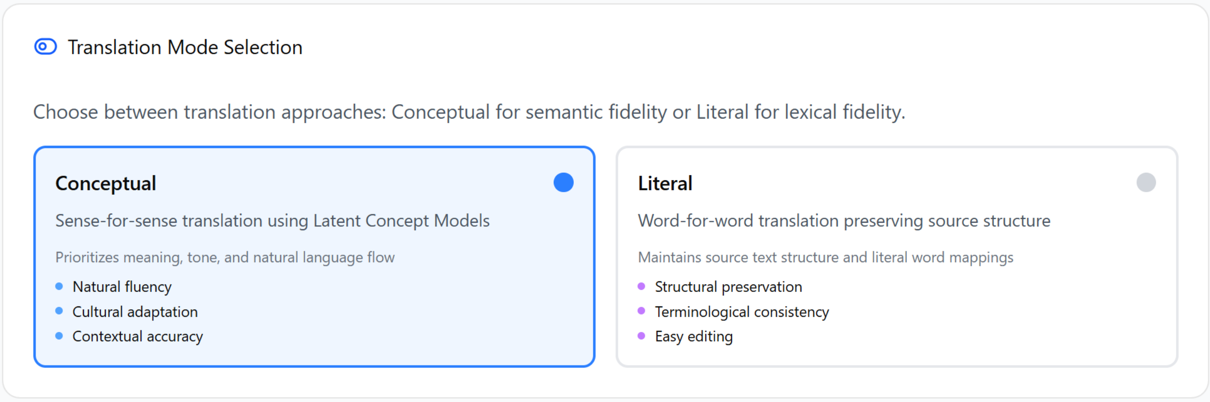
When applying these new models, we can also consider how humans approach translation and how they represent concepts across languages. Findings from cognitive science can help identify which translation contexts benefit most from different approaches, and which interface features might reduce cognitive load for post-editors. Cognitive principles can also guide the development of more human-aligned evaluation metrics, making both post-editing and system scoring more robust. To make MT more natural and human-like, much can be learned by analyzing where these systems align with and where they do not align with human cognition.

2 Future Directions for MTPE

2.1 Balancing conceptual and lexical accuracy

LCMs differ from traditional LMs by predicting the next concept rather than the next token in a sequence. This approach has the potential to improve translations by prioritizing the text's abstract meaning over matching the most probable word sequence. Human translators and interpreters are often described as operating along a spectrum from word-for-word (literal) and sense-for-sense (free) translation (Blanchot, 1990). Word-for-word translation aims to preserve the vocabulary and grammatical structure of the source text as much as possible in the target language. Meanwhile, sense-for-sense translation focuses on conveying the meaning and tone of the source text naturally in the target language. Although the balance between preserving form and conveying message is subjective and context-dependent, LCMs' concept-based representations may reduce PE effort by aligning more closely with free translation strategies. They may also support new evaluation metrics that assess semantic fidelity rather than surface-level string similarity.

Traditional MT systems are more likely to struggle with LRLs because they often lack sufficient high-quality training data to produce robust translations. As a result, LRL translation tends to require more extensive PE. This raises an essential question for MTPE: is it more cognitively demanding to edit a literal, word-for-word translation that misses intended meaning, or a looser, sense-for-sense translation that sacrifices lexical fidelity? LCMs allow us to empirically test this question because they are designed to capture higher-level concepts, whereas traditional token-based LMs focus on word patterns. In particular, experiments could test the specific advantages they might confer for LRLs or distant language pairs. Such experiments could compare the time, effort, and preferences of editors when correcting concept-based versus token-based translations, across both high- and low-resource language pairs. One possible outcome of this research is that the preferred model depends on the text or the editor. In this case, interfaces could be adapted to support toggling between concept-aligned and token-aligned views, as illustrated in Figures 1 and 2, helping editors decide when fidelity to the source or fluency in the target language



Translation Mode Selection

Choose between translation approaches: Conceptual for semantic fidelity or Literal for lexical fidelity.

Conceptual ●

Sense-for-sense translation using Latent Concept Models

Prioritizes meaning, tone, and natural language flow

- Natural fluency
- Cultural adaptation
- Contextual accuracy

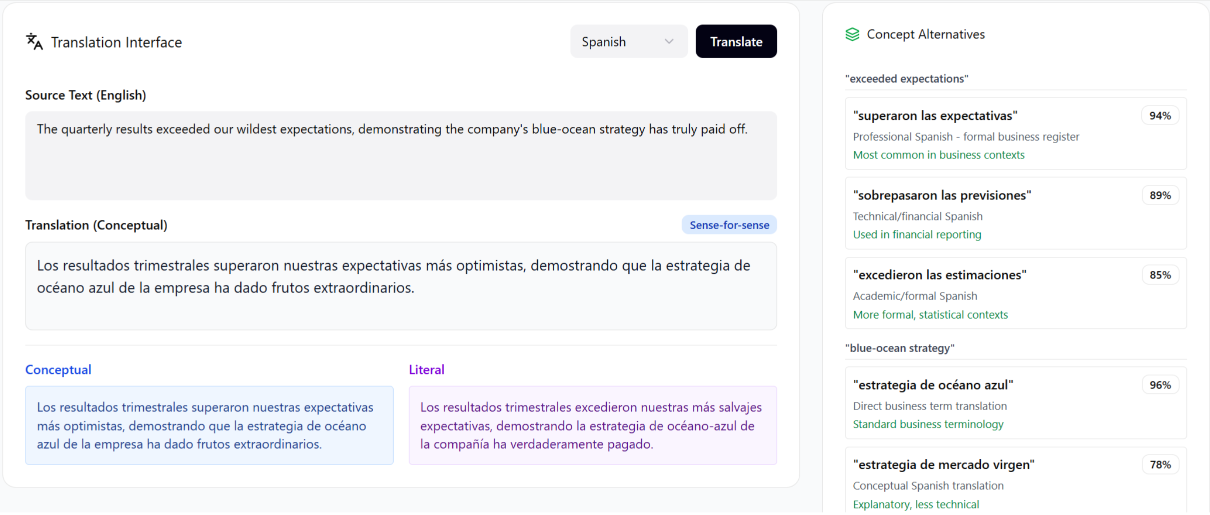
Literal ○

Word-for-word translation preserving source structure

Maintains source text structure and literal word mappings

- Structural preservation
- Terminological consistency
- Easy editing

Figure 1: A mockup of an MTPE interface feature allowing editors to choose between conceptual and literal translation modes.



Translation Interface

Source Text (English)

The quarterly results exceeded our wildest expectations, demonstrating the company's blue-ocean strategy has truly paid off.

Translation (Conceptual) — Sense-for-sense

Los resultados trimestrales superaron nuestras expectativas más optimistas, demostrando que la estrategia de océano azul de la empresa ha dado frutos extraordinarios.

Conceptual

Los resultados trimestrales superaron nuestras expectativas más optimistas, demostrando que la estrategia de océano azul de la empresa ha dado frutos extraordinarios.

Literal

Los resultados trimestrales excedieron nuestras más salvajes expectativas, demostrando la estrategia de océano-azul de la compañía ha verdaderamente pagado.

Concept Alternatives

"exceeded expectations"

"superaron las expectativas" 94%
Professional Spanish - formal business register
Most common in business contexts

"sobrepasaron las previsiones" 89%
Technical/financial Spanish
Used in financial reporting

"excedieron las estimaciones" 85%
Academic/formal Spanish
More formal, statistical contexts

"blue-ocean strategy"

"estrategia de océano azul" 96%
Direct business term translation
Standard business terminology

"estrategia de mercado virgen" 78%
Conceptual Spanish translation
Explanatory, less technical

Figure 2: An example translation interface layout, including side-by-side comparison of translation approaches (Section 2.1) and a module displaying candidate translations for concepts (Section 2.2), evaluated by their fit with the current context.

is more important.

2.2 Language, cognition, and cross-linguistic representations

Although the LCM was intended to be more "human-like" by using abstract, language-agnostic representations, research shows that semantic spaces (i.e., the way meaning is structured and related in memory or model embeddings) depend partly on the language being used (Chen et al., 2024; Zada et al., 2025). For example, Greek has different categories to represent what would be labeled "blue" in English. Greek-English bilinguals' representations of color concepts shift depending on language context: the more dominant their Greek use, the more distinctly they separate categories such as "ghalazio" (light blue) and "ble" (dark blue); with stronger English dominance, these categories merge more closely (Athanasopoulos, 2009). Similarly, Mandarin-English bilinguals will automatically retrieve different answers to the prompt "Name a statue of someone standing with a raised arm while looking into the distance" when asked in Mandarin, where they say the Statue of Mao, versus English, where they say the Statue of Liberty. The way concepts are represented in human memory can shift significantly depending on the language context.

Therefore, a more human-like cross-linguistic model would retain the LCM's abstraction capabilities. Still, rather than aiming to be wholly language-independent, it can adapt to the language of the text it is processing or producing. Additionally, evidence that language shapes concepts raises questions about whether LCMs can ever truly achieve language-agnosticism. Wu et al. (2025) demonstrated that LLMs have a shared multilingual semantic representation space, but it is “anchored” to the dominant languages of the model’s training data. In other words, if an LLM is trained primarily on English, its embeddings will be biased towards the conceptual structure of English, even when performing tasks in other languages. Thus, even if LCMs aim to encode universal conceptual embeddings, training on an uneven distribution of languages may bias the semantic space toward the conceptual structures of dominant languages. Before employing LCM-like models as tools for LRL MT and MTPE, a key line of research will be to thoroughly test whether their predominant languages scaffold them, as LLMs and humans are.

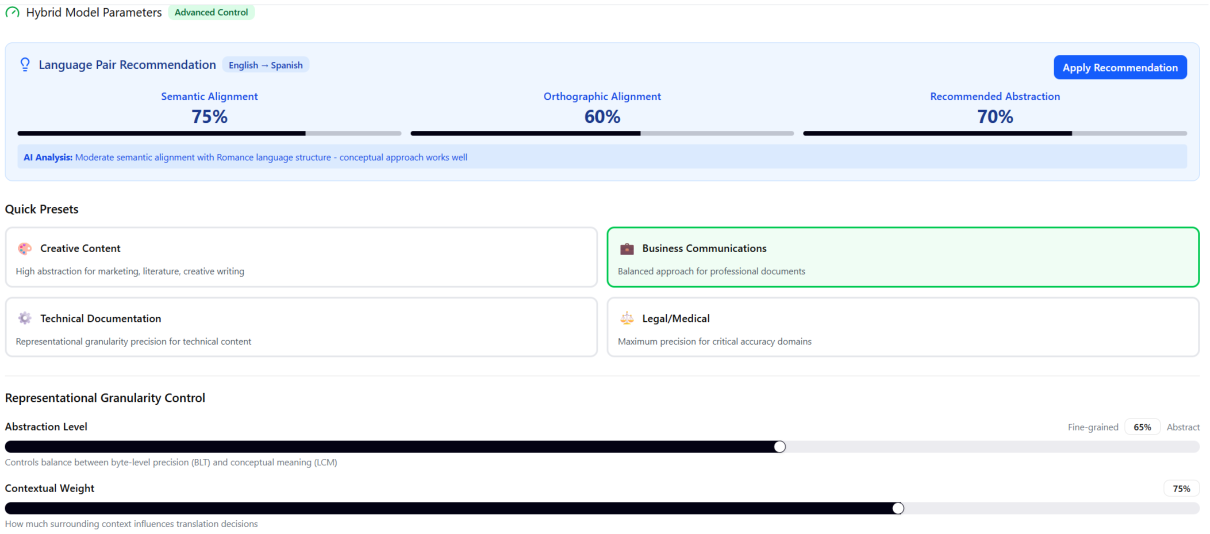
If LCMs can capture conceptual spaces across languages, they may enable more flexibility in the translations given to post-editors. For instance, when translating “blue” from English to Greek, the system could recognize that multiple potential Greek translations overlap with that concept and offer a ranked set of candidate translations to the post-editor. One advantage of broad, abstract representations is that they can map flexibly onto multiple concrete linguistic expressions. This feature could be leveraged in MTPE interfaces by presenting editors with multiple translation options for ambiguous concepts, allowing them to select the most contextually appropriate form (see Figure 2). Interfaces could log editors’ choices, generating valuable data to improve MT in low-resource contexts.

2.3 Optimal uses for LCM and BLT approaches in MT

The LCM and BLT represent contrasting approaches to semantic representation. The former is based on principles of abstraction, whereby concepts are encoded into generalized representations that are invariant to specifics of the context. The latter encodes text at the byte level, dynamically segmenting character sequences. This makes its representations more fine-grained and context-sensitive than those of token-based models or LCMs. For example, a sentence like “On June 1st, we spent several hours sitting in the dewy grass of Central Park, enjoying the sunshine” might, in an LCM-like model, be abstracted into the overarching concept of an afternoon in the park, while still retaining information about participants and actions. By contrast, a BLT-like model would process the sentence by dynamically segmenting its byte stream, encoding information at the level of each character sequence.

Our memory system relies on both types of mechanisms to effectively store and organize information. For instance, it might not be necessary to store all the contextual details about the experience at the park, so categorizing it under the more abstract concept of ’afternoon in the park’ is more efficient and allows one to integrate prior knowledge about this type of event to generalize and make relevant inferences. From this general label and our understanding of parks, we can fill in gaps and infer that the experience was outdoors, likely with nice weather, and included typical park features, such as grassy areas or benches. However, in some circumstances, a conceptual representation must be tied to the specific context in which it was experienced. Key information about the experience might be dependent on the particular date (e.g., a birthday) or location (e.g., Central Park) where it occurred. People rely more on abstract or context-specific representations depending on task demands (Barsalou, 1999; Yee and Thompson-Schill, 2016). Consequently, a more human-like system might integrate both approaches, flexibly shifting between or combining abstract concept-based embeddings and fine-grained byte-level representations depending on the translation context. Hybrid LCM/BLT outputs could allow post-editors to toggle between the two. This choice may relate to whether the word-for-word or sense-for-sense approach described earlier is better suited for the domain at hand. In addition, the message’s intent may dictate whether a more abstract or a more concrete representation is more effective for translation.

The effectiveness of either model type in MT may also depend on the language pair. Thompson et al. (2020) analyzed semantic alignment across 41 languages, defined by the degree to which equivalent words across two languages occupy similar positions in the semantic embedding space relative to other words. They found that the degree of semantic alignment between a pair of languages was predicted by their historical, geographic, and cultural relatedness. In other words, languages with similar geographic and cultural backgrounds organize the world into similar concepts through the words they have to label them. Because the LCM assumes that all languages share a universal conceptual embedding space, its translation success across a given pair of languages may be predicted



Hybrid Model Parameters

Language Pair Recommendation English → Spanish

Semantic Alignment 75%

Orthographic Alignment 60%

Recommended Abstraction 70%

AI Analysis: Moderate semantic alignment with Romance language structure - conceptual approach works well

Quick Presets

Creative Content
High abstraction for marketing, literature, creative writing

Business Communications
Balanced approach for professional documents

Technical Documentation
Representational granularity precision for technical content

Legal/Medical
Maximum precision for critical accuracy domains

Representational Granularity Control

Abstraction Level — Fine-grained 65% Abstract
Controls balance between byte-level precision (BLT) and conceptual meaning (LCM)

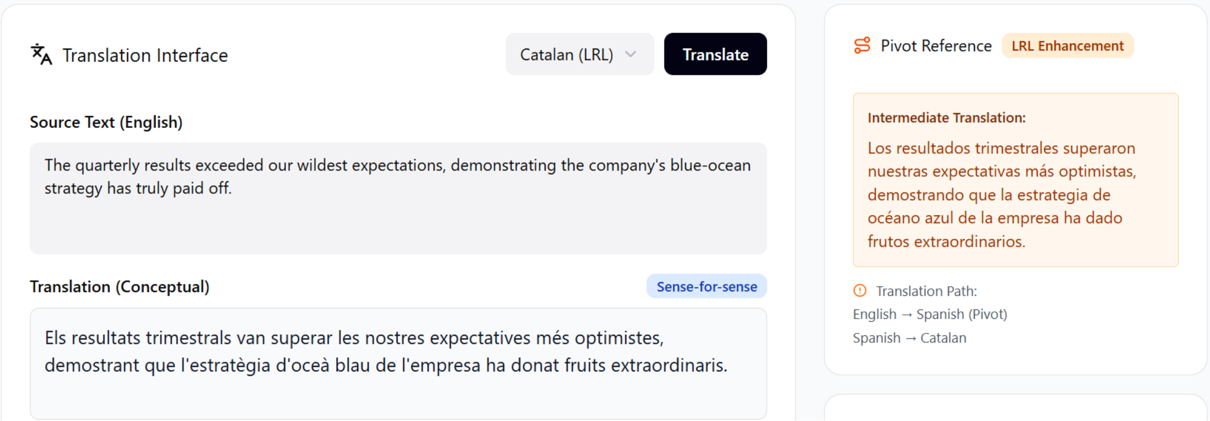
Contextual Weight 75%
How much surrounding context influences translation decisions

Figure 3: An example hybrid model control panel that recommends optimal abstraction settings based on the language pair selected and offers preset options for use cases requiring more abstract or more precise translations.

by the degree of semantic alignment and by their historical, geographic, and cultural similarity. Future research can test this hypothesis by evaluating LCM-based MT on language pairs that vary across these factors. If such a relationship holds, it would suggest that LCM-based translation is especially advantageous for specific language pairings and could guide decisions about when to deploy LCMs versus traditional MT models.

Conversely, BLTs are tuned to representations at the byte level (i.e., the raw encoded symbols and characters of a language) rather than abstract conceptual mappings. This focal point of the model architecture presents a parallel research question to the previous one: do BLT-based translation systems perform better with orthographically similar language pairs? Orthography refers to the written component of a language, including its characters, spelling, capitalization and punctuation norms, all of which become the basis for embeddings in a BLT. Although BLTs were promoted as promising for LRL translation, their byte-level representations may actually favor language pairs with shared orthographic features, since similar scripts and character sets reduce the complexity of cross-lingual alignment. For example, languages that share an alphabet, like English and Italian, might yield better results than English and Chinese, which use different sets of symbols that carry different amounts of information per unit. Experiments could systematically compare BLT performance across language pairs with varying degrees of orthographic similarity (e.g., shared alphabet vs. distinct scripts) to assess whether byte-level sensitivity offers measurable advantages in editing speed or accuracy. The findings could inform when and how BLTs are applied in the MTPE process. Figure 3 shows an example of how editors could adjust the settings of a hypothetical hybrid model based on recommendations about language pair alignment.

Taken together, understanding the types of language pairs where different models excel could also aid LRL translation by identifying optimal paths for indirect translation when direct translation is difficult, also called pivoting. LRLs could be paired with a higher-resource "pivot" language that is either conceptually or orthographically closely aligned. When translating to or from the LRL, an initial translation could be made into its "pivot" language using a generic MT model before using a more specialized LCM or BLT model for the final translation. For example, Catalan could be translated into semantically similar Spanish before reaching English. Pivot-based strategies have long been used to overcome the challenges of LRL translation and evaluation (Paul et al., 2013; Mukherjee et al., 2025; Lakew et al., 2017), though they rarely leverage different model architectures across translation steps. Even if not integrated directly into the main MT pipeline, the intermediate outputs could be presented alongside the draft translation in the MTPE interface, giving editors additional reference points that may reduce search effort and improve efficiency (see Figure 4).



Source Text (English)

The quarterly results exceeded our wildest expectations, demonstrating the company's blue-ocean strategy has truly paid off.

Translation (Conceptual)

Els resultats trimestrals van superar les nostres expectatives més optimistes, demostrant que l'estratègia d'oceà blau de l'empresa ha donat fruits extraordinaris.

Intermediate Translation:

Los resultados trimestrales superaron nuestras expectativas más optimistas, demostrando que la estrategia de océano azul de la empresa ha dado frutos extraordinarios.

Translation Path:
English → Spanish (Pivot)
Spanish → Catalan

Figure 4: When translating to or from an LRL, the MTPE interface could provide a reliable pivot reference in a related language to reduce effort for editors evaluating the MT output.

2.4 Leveraging human PE evaluation data to improve MT in non-traditional models

Human annotation data, such as post-edit corrections and error labels, has been shown to effectively improve MT in LLMs through a variety of techniques (e.g., Ki and Carpuat, 2024; Koneru et al., 2024; Raunak et al., 2023), with recent work demonstrating particular promise for LRLs (Deoghare et al., 2024). This type of data is more valuable than simple reference translations because it shows human strategies for correcting actual errors made by MT systems. As emerging architectures such as LCMs and BLTs continue to develop, MTPE data may offer similar benefits by aligning their outputs with human translation practices. Through human error corrections and fine-tuning, an LCM translation system could refine its conceptual embeddings and compensate for shortcomings such as missing lexical coverage or mismatched cultural associations, thereby aligning more closely with human expectations than a purely distributional model. Likewise, MTPE data could strengthen BLTs by guiding them towards more effective mappings between orthographic forms and meaning, particularly when byte-level representations alone fail to capture semantic nuance. While neither LCMs nor BLTs can fully replicate the cognitive processes involved in human translation, MTPE feedback provides a practical mechanism for approximating them. Incorporating insights from such data into system design not only improves translation accuracy but also allows interfaces to highlight common error types and adapt to individual editor preferences.

3 Discussion

Recent moves away from token-based LLMs raise new theoretical questions and present opportunities to redesign MTPE workflows and interfaces, especially with respect to the unique challenges posed by LRLs. In this paper, we focus on the Large Concept Model and the Byte Latent Transformer and examine several topics in light of relevant cognitive scientific theories. We also analyze their implications for future research and design in MTPE, summarized below:

1. LCMs may produce translations that prioritize the meaning of the source text over word-for-word accuracy, thereby reducing PE effort by aligning more closely with human translation strategies than traditional MT models. This potential improvement could be particularly apparent for LRLs, whose MTs are more likely to suffer in quality due to lack of training data.

2. While the LCM can generate text in LRLs better than LLMs, its current embedding model was only trained on English, which may bias the learned concept space and distort cross-lingual mappings, limiting effectiveness of non-English pairs. This should be tested to guide future assumptions about the appropriate use of LCMs in MT and MTPE.

3. LCM-like architectures could be used to offer post-editors multiple translation options for ambiguous texts.

4. An ideal hybrid LCM-BLT system would dynamically adjust the granularity of its seman-

tic representations based on task context, producing more human-like MTs and reducing PE effort.

5. LCM translation quality may depend on the degree of semantic overlap between the languages, while BLT quality may be more sensitive to orthographic similarity. These patterns could help determine when each model should be used inthe translation workflow.
   (a) If either model shows sensitivity to these linguistic relationships, an LRL MT could potentially be improved with an intermediary translation through a higher-resource language that is well-matched in semantic structure or orthography.

6. Human MTPE data can help tune both LCMs and BLTs to improve translation capabilities.

LCMs and BLTs each address the limitations of token-based LMs in promising ways, one through representations above the level of individual tokens and the other through representations below the level of individual tokens. While each has the potential to advance LRL translation and MTPE, it is critical to consider the assumptions underlying any model and their implications. Research on the human mind can help generate hypotheses about the conditions under which models will perform well and how best to facilitate human-in-the-loop work. By integrating interdisciplinary insights, we can continue to maximize the benefits of MTPE, ensuring more equitable access to reliable translation technology across languages, including those with fewer resources.

References

Panos Athanasopoulos. 2009. Cognitive representation of colour in bilinguals: The case of greek blues. *Bilingualism: Language and cognition*, 12(1):83–95.

Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, David Dale, and 1 others. 2024. Large concept models: Language modeling in a sentence representation space. *CoRR*.

Lawrence W Barsalou. 1999. Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660.

Maurice Blanchot. 1990. Translating. *Sulfur*, (26):82.

Catherine Chen, Xue L Gong, Christine Tseng, Daniel L Klein, Jack L Gallant, and Fatma Deniz. 2024. Bilingual language processing relies on shared semantic representations that are modulated by each language. *bioRxiv*, pages 2024–06.

Sourabh Deoghare, Diptesh Kanojia, and Pushpak Bhattacharyya. 2024. Together we can: Multilingual automatic post-editing for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10800–10812.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Dayeon Ki and Marine Carpuat. 2024. Guiding large language models to post-edit machine translation with error annotations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4253–4273.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024. Contextual refinement of translations: Large language models for sentence and document-level post-editing. In *NAACL-HLT*.

Surafel M Lakew, Quintino F Lotito, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Improving zero-shot translation of low-resource languages. In *Proceedings of the 14th International Workshop on Spoken Language Translation*.

Ananya Mukherjee, Saumitra Yadav, and Manish Shrivastava. 2025. Why should only high-resource-languages have all the fun? pivot based evaluation in low resource setting. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4779–4788.

Artidoro Pagnoni, Ram Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason Weston, Luke Zettlemoyer, and 1 others. 2024. Byte latent transformer: Patches scale better than tokens. *arXiv preprint arXiv:2412.09871*.

Michael Paul, Andrew Finch, and Eiichrio Sumita. 2013. How to choose the best pivot language for automatic translation of low-resource languages. *ACM Transactions on Asian Language Information Processing (TALIP)*, 12(4):1–17.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024.

Bill Thompson, Seán G Roberts, and Gary Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10):1029–1038.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations*.

Eiling Yee and Sharon L Thompson-Schill. 2016. Putting concepts into context. *Psychonomic bulletin & review*, 23(4):1015–1027.

Zaid Zada, Samuel A Nastase, Jixing Li, and Uri Hasson. 2025. Brains and language models converge on a shared conceptual space across different languages. *arXiv preprint arXiv:2506.20489*.